Aplicaciones básicas en la Red para la explotación y representación de datos

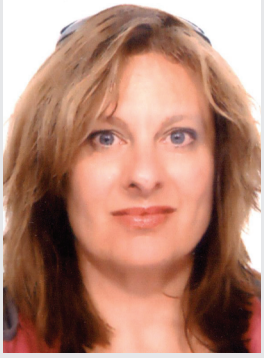
Basic data analysis and presentation tools on the Net

Luis-Millán González-Moreno y Fernanda Peset

Gónzález-Moreno, Luis-Millán; **Peset, Fernanda** (2015). "Aplicaciones básicas en la Red para la explotación y representación de datos". *Anuario ThinkEPI*, v. 9, pp. 254-259.

http://dx.doi.org/10.3145/thinkepi.2015.60

Publicado en *IweTel* el 10 de febrero de 2015



Resumen: Se reflexiona sobre la necesidad de evolucionar en el campo de los *open data* desde la etapa de la producción hasta su explotación. Se detalla el proceso para generar una visualización simple: conocimiento de las fuentes y extracción; tratamiento en local o en la nube; manejo de las herramientas necesarias para representarlos. Se ofrece información sobre algunas utilidades gratuitas de la Red que un documentalista debería conocer. Por último, se reinterpreta el proceso de adquisición de conocimiento en el campo de datos ilustrándolo con un ejemplo que ofrece visualizaciones avanzadas basadas en datos abiertos: *London: the information capital*.

Palabras clave: Datos abiertos; Minería de datos; Explotación de datos; Visualización de datos; Aplicaciones.

Abstract: We offer some thoughts on the need for the field of open data to evolve from the production to exploitation stage. We detail the process of generating a simple display: knowledge of sources and data extraction, local analysis or in the cloud, and using the tools needed to present the data. Information is offered on free tools readily available on the Net that an information specialist should know. Finally, knowledge acquisition in the field of data is reinterpreted, and illustrated with an example, *London: the information capital*, that provides advanced visualizations based on open data.

Keywords: Open data; Data mining; Data visualization; Applications.

Ingentes cantidades de datos

Desde hace un tiempo escuchamos la frase, atribuida a **Neelie Kroes** (2012): los datos van a ser el nuevo combustible/oro/motor/petróleo de la economía. Esto entra de lleno en la órbita de los intereses de quienes gestionamos la información. Vocablo éste que recientemente ha evolucionado un poco y que en ocasiones vemos sustituido por datos. Este avance ha tenido que ver con otros ecosistemas que influyen en el trabajo con información a escala mundial: desde los *linked data* a los *big data*, pasando, obviamente por *open data*.

Después de unos años observando e investigando estos temas cabe preguntarse si los datos por sí solos producirán una bonanza económica, tal como algunos prometen. De hecho, se detectan elipsis en los planteamientos sobre *open data* que ya fueron analizadas por varios autores (**Ferrer-Sapena**; **Sánchez-Pérez**, 2013; **Bates**, 2013). Se da como cierto que abrir los datos producirá una mejora en toda la economía (**Abella**; **Ortiz-de-Urbina-Criado**; **De-Pablos-Heredero**, 2014). Pero este salto es una presunción. Todavía no contamos con indicadores que lo demuestren fehacientemente.

"No es suficiente con hacer disponibles los datos"

Figura 1. Representación en mapa de la cantidad de bancos de datos registrados en *Databib* por país. Realizado en *CartoDB*.
http://cdb.io/19zsglJ

Pero para que sean útiles hay que hacerlos asimilables

Consideramos que no es suficiente con hacer disponibles los datos, razón por la que este texto plantea la necesidad de explorar unas mínimas técnicas de explotación que puedan hacer productivas esas prometedoras masas de información. De forma adicional, esto hará dar un paso adelante al sector de la gestión de la información. Se trata de un sector acostumbrado a hacer disponible la información de forma interoperable, semántica, abierta, normalizada, etc. Pero esto ahora no es ya suficiente. El bibliotecario integrado (**Torres-Salinas**, 2011) debe estar en línea con los intereses de los investigadores. En palabras de **Monastersky** (2013) debe ascender en la cadena de producción de la ciencia.

Este texto recomienda aplicaciones más o menos simples que acometen la explotación de los datos, así como los textos en las que un documentalista debería formarse. Las que recogemos se encuentran en la Red de forma gratuita; por ejemplo, las nubes de *tags* que, dentro de su simplicidad, siguen siendo eficientes para representar la frecuencia de aparición de un término en un conjunto de palabras. Aunque también resulta útil conocer técnicas derivadas de la bibliometría que explotan los datos, ya sea utilizando las herramientas de análisis de las propias bases de datos bibliográficas (*Scopus* o *Web of science*), ya sea combinando *Bibexcel* y *Pajek* para generar redes de *co-words* (**García-García** *et al.*, 2015).

En suma, consideramos que un profesional de la información debe manejar mínimamente herramientas como *Many eyes*, *Tableau* o *CartoDB*. Una vez dicho esto, que no es poco, este thinkepi intenta ir paso a paso mostrando algunos ejemplos, dado el valor de la explotación de datos y textos con métodos automáticos (**McDonald**; **Kelly**, 2012). El proceso completo para llegar a generar una figura o una visualización constaría de tres fases:

Figura 2. Representación del *Grupo Thinkepi* en burbujas, por lugar de procedencia (realizado con *Tableau*).
https://public.tableausoftware.com/views/MiembrosThinkepiporLugar/Sheet1

- aplicar técnicas de extracción de información (no hablamos de generar nuevos datos), ya sea exportando registros bibliográficos, ya sea localizando fuentes abiertas, como serían para el campo de la información *DOAJ*, *SJR* o *DataBib*;
- contar con habilidades para el análisis y manejo de ficheros de datos estructurados: desde limpiar los ficheros, realizar análisis descriptivos, a migrar formatos o subirlos a *Drive*, *Dropbox*…
- conocer las herramientas necesarias para representar gráficamente una idea.

"Las técnicas de explotación de datos pueden hacer productivas esas prometedoras masas de información"

Extracción de información

Puede ser costosa y precisar de una gran capacidad de computación si se trabaja con grandes masas de texto en lenguaje natural y no estructurado. Por ejemplo, un análisis sobre las políticas de revistas con respecto al material adicional (**García-García** *et al.*, 2014) precisó un ingente trabajo manual de preparación de los textos para aplicar los algoritmos de relación. Esto desaconsejó usarlo como método automático de análisis.

Sin embargo se pueden usar fácilmente fuentes de información más o menos estructurada y abierta. **Delgado** (2014) ofrece un listado entre los que menciona boletines oficiales (*BOE* o *Borme*), institutos oficiales de estadística (*INE* o *Eurostat*), fuentes con información bancaria (*Banco de España*, *World Bank* o *FMI*), etc.

Asimismo tampoco hay que olvidar algunos desarrollos concretos como *GapMinder* (**Rosling**, 2010), que proporcionan a un tiempo los datos y su visualización. La posibilidad creciente de contar con datos y textos abiertos, es decir que admitan su reutilización para fines distintos para los que fueron concebidos, hace prever una eclosión en su explotación (**Murray-Rust**; **Molloy**; **Cabell**, 2014).

Técnicas de explotación de datos

La aplicación de este tipo de técnicas está avanzando a pasos agigantados. De hecho, el think tank ubicado en Bruselas *Lisbon Council for Economic Competitiveness and Social Renewal* solicitó un informe sobre la investigación europea en *data y text mining* (**Filippov**, 2014). En él se ofrece un panorama descriptivo de la situación y sus problemas, aunque no describe las técnicas que pueden aplicarse. Para conocerlas, **Berger** (2012) explica las que pueden ejecutarse con *Oracle*: clasificación, regresión, detección de anomalías, importancia de atributos, *clustering* (agrupación), extracción de características… (**Serrano-Cobos**, 2014). Pero otras, también avanzadas como por ejemplo los análisis de redes neurales, pueden realizarse con *MatLab* (**García-García** *et al.*, 2013).

Programas para la representación gráfica

Por último, la utilización de las herramientas citadas puede dar lugar presentaciones más o menos acertadas. En estos momentos tenemos a nuestro alcance múltiples programas de visualización (**Nualart**; **Pérez-Montoro**; **Whitelaw**, 2014) capaces de aplicar técnicas de análisis y

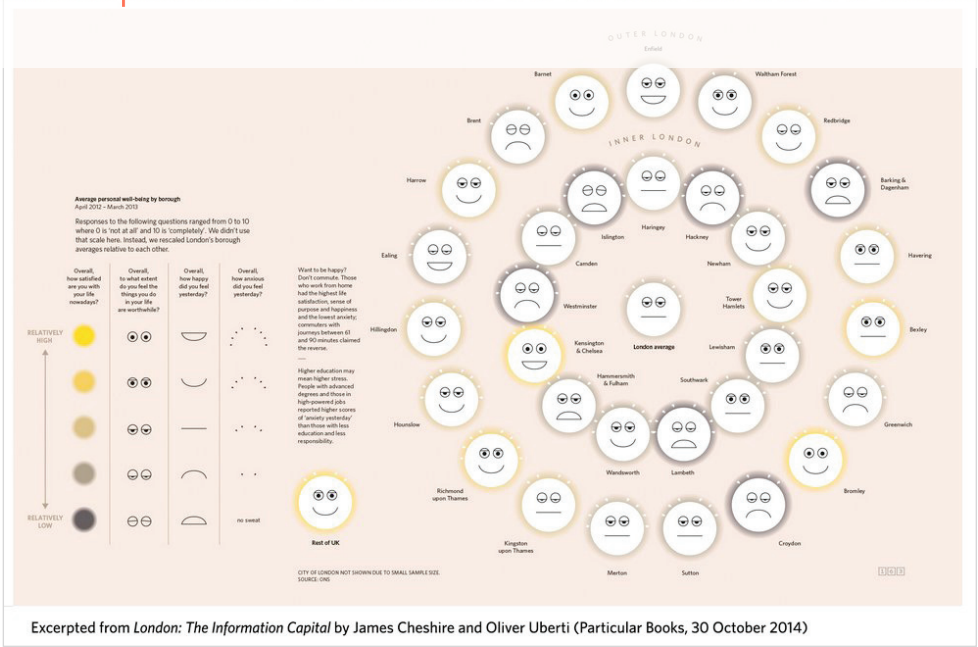
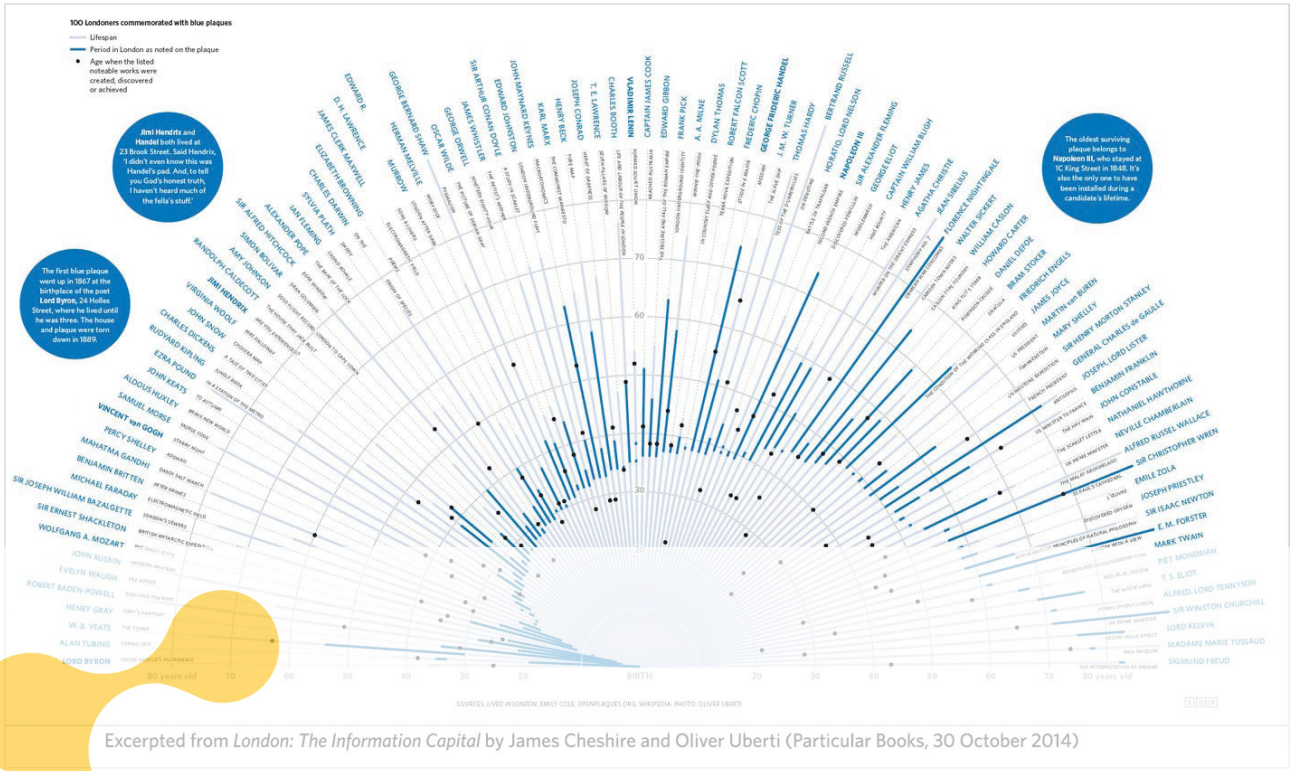


Figura 3. "*Islington has issues. New research seeks to improve well-being by identifying who's hurting*" (**Cheshire**; **Uberti,** 2014, p. 163). Fuente de los datos: *Office for National Statistics*.
http://theinformationcapital.com/project/islington-has-issues



Excerpted from *London: The Information Capital* by James Cheshire and Oliver Uberti (Particular Books, 30 October 2014)

Figura 4. *Who London inspired (from Mozart to Lenin, blue plaques commemorate those who called the city home).* En el gráfico se representan 100 londinenses famosos, que tienen placas de color azul en los edificios donde residieron. La línea gris es su vida, la azul el tiempo que vivieron en la ciudad y los puntos la edad que tenían cuando se hicieron famosos (**Cheshire**; **Uberti**, 2014).
http://theinformationcapital.com/project/who-london-inspired

dibujar vistosas presentaciones para web de forma bastante intuitiva:

- *Tableau*, lo podemos ver utilizado en *Scholarly Publishers Index* (*SPI*): *http://get.tableau.com/es-es/trial/tableau-software.html*
- *Many eyes*, es un potente software de *IBM*: *http://www-01.ibm.com/software/analytics/many-eyes*
- *CartoDB*. es una aplicación de fácil uso, con versiones gratuita y premium. La empresa tiene oficinas en Madrid y New York. *http://cartodb.com*

Todos funcionan de manera similar en su versión gratuita por web, aunque el primero cuenta con una aplicación de escritorio que es necesario descargar. El procedimiento es importar datos estructurados y escoger el estilo de gráfico que más se adecue a la idea que queremos representar. A modo de ejemplo, ofrecemos la frecuencia del lugar de procedencia del *Grupo Thinkepi* (figura2).

Representaciones más impactantes aún pueden también hacerse con otros softwares gratuitos. **Torres-Salinas** (2014) muestra una animación de la evolución de la investigación española en biomedicina basándose en la hoja de cálculo de *Google docs*.

De los datos a la sabiduría

Como reflexión final querríamos plantear un símil con aquella gradación que todos conocemos que otorga un valor creciente a datos, información, conocimiento y, por último, saber (**Rowley**, 2007):

- el primer escalón, abrir los datos haciéndolos disponibles, constituiría la base necesaria para que puedan utilizarse para otros fines;
- el segundo, la información, estaría compuesto por los resultados de explotar de manera aislada cada uno de los *datasets* disponibles;
- en la tercera fase, esos análisis recombinarían y utilizarían los recursos que puedan resultar eficientes para esa explotación, innovando con nuevo conocimiento;
- por último, el saber se cifraría en devolver a la sociedad esos nuevos *datasets* y visualizaciones de manera que se pongan en valor todas esas costosas, complejas y sofisticadas operaciones que han sido necesarias para llegar al estadio más avanzado.

Como sabemos que es difícil imaginar este camino en abstracto vamos a utilizar un sencillo ejemplo para ilustrar esa gradación. Tomemos la cuidada edición de **Cheshire** y **Uberti** (2014) del libro *London: the information capital: 100 maps and graphics that will change how you*

view the city en la que, a nuestro entender, se pueden observar todas estas etapas. Los autores han contado con datos liberados, en buena parte procedentes del gobierno inglés tras la reunión de Tim Berners-Lee con el primer ministro Gordon Brown en 2009. En ese primer momento identifican los que pueden ser de interés entre una numerosa masa de datos; datos libres pero inertes. En una segunda etapa, empiezan a darles vida, analizándolos con herramientas sofisticadas para representarlos de manera atractiva para su impresión (figura 3). Pero por muy vistosa que sea la visualización, no producen una especial innovación ya que hacen uso de un solo *dataset*.

Podemos hablar de un tercer nivel en el que se crea nuevo conocimiento cuando se combinan varias fuentes de datos en nuevas visualizaciones. En esa etapa se han tenido que localizar y armonizar las fuentes y reflexionar de forma creativa qué idea se quiere transmitir y cómo. Este avance en el proceso de conocimiento es sutil pero esencial, y requiere capacidades más allá del manejo de las técnicas y métodos de los que hablábamos al comienzo de este texto. La figura 4, por ejemplo, utiliza tres fuentes: *Lived in London*, de Emily Cole; *Openplaques.org*; y *Wikipedia* para crear la visualización.

Y por último la cima del proceso, el saber, trasciende al propio conocimiento. Algo que se produce en pocas ocasiones. En nuestra consideración, el saber es la utilización inteligente del conocimiento y su puesta a disposición para el disfrute de la sociedad. Es una puesta en valor del conocimiento, no hay que confundir, se cifra exactamente en parámetros de explotación económicos. En el ejemplo que presentamos, sus autores han puesto a disposición del ciudadano una publicación que permite gozar de la belleza que deriva de la explotación de los datos. Un deleite que procede de su naturaleza artística, lo que constituye para **Freud** (1929) una de las escasas satisfacciones con las que cuenta el hombre para aliviar su "sufrimiento".

Notas

1. Agradecemos a Israel Pedrós su conocimiento de *Tableau* y otras herramientas.

2. Declaración de conflicto de intereses, si procede: Fernanda Peset es directora de un subproyecto coordinado del Plan Nacional I+D+i sobre Gestión de datos de investigación y tiene interés en que se hable cuanto más sea posible de datos en todas sus formas ☺.

Bibliografía

Abella, Alberto; **Ortiz-de-Urbina-Criado, Marta**; **De-Pablos-Heredero, Carmen** (2014). "*Meloda*, métrica para evaluar la reutilización de datos abiertos". *El profesional de la información*, v. 23, n. 6, pp. 582-588. *http://www.elprofesionaldelainformacion.com/contenidos/2014/nov/04.pdf*
http://dx.doi.org/10.3145/epi.2014.nov.04

Bates, Jo (2013). "Opening up public data". *Speri.comment: the political economy blog*, 21 May. *http://speri.dept.shef.ac.uk/2013/05/21/opening-public-data*

Berger, Charlie (2012). *Big data analytics with Oracle advanced analytics in-database option*. *http://goo.gl/CgxYAL*

Cheshire, James; Uberti, Oliver (2014). *London: the information capital: 100 maps and graphics that will change how you view the city*. Particular books. *http://theinformationcapital.com*

Delgado, Antonio (2014). "Periodismo de datos: una introducción a esta disciplina del periodismo". En *1st Intl workshop on open research data*. *http://mugi.webs.upv.es/1st-international-workshop-open-research-data*

Ferrer-Sapena, Antonia; **Sánchez-Pérez, Enrique-A.** (2013). "Open data, big data: ¿hacia dónde nos dirigimos?". *Anuario Thinkepi*, v. 7, pp. 150-156. *http://eprints.rclis.org/21006*

Filippov, Sergey (2014). *"Mapping text and data mining in academic and research communities in Europe". Lisbon Council 2014*, special briefing, n. 16. *http://goo.gl/330Jtz*

Freud, Sigmund [1929]. *El malestar en la cultura*. *http://www.dfpd.edu.uy/ifd/rochalm_apoyo/2/sig_freud_el_malestar_cult.pdf*

García-García, Alicia; **García-Massó, Xavier**; **Ferrer-Sapena, Antonia**; **González-Moreno, Luis-Millán**; **Fernanda Peset** (2013). "¿Es reutilizable el material adicional en las revistas más citadas?". En: *3ª Conf intl sobre revistas en ciencias sociales y humanidades (CRECS 2013)*, Sevilla, 9 mayo. *http://eprints.rclis.org/21001*

García-García, Alicia; **García-Massó, Xavier**; **Ferrer-Sapena, Antonia**; **González-Moreno, Luis-Millán**; **Peset, Fernanda**; **Villamón-Herrera Miguel**; **Aleixandre-Benavent, Rafael** (2014). "Text mining versus redes neuronales. Dos métodos de análisis aplicados al caso de las políticas de las revistas sobre datos". En: *4ª Conf intl sobre calidad de revistas de ciencias sociales y humanidades* (*CRECS 2014*), Madrid, 8-9 de mayo.

García-García, Alicia; **Pardo-Ibáñez, Alberto**; **Ferrer-Sapena, Antonia**; **Peset, Fernanda**; **González-Moreno, Luis-Millán** (2015, en prensa). "Herramientas de análisis de datos bibliográficos y construcción de mapas de conocimiento: Bibexcel y Pajek". *BiD: textos universitaris de biblioteconomia i documentació*, junio.

McDonald, Diane; **Kelly, Ursula** (2012). *The value and benefit of text mining to UK further and higher education. Digital infrastructure*. JISC. *http://bit.ly/jisc-textm*
http://www.jisc.ac.uk/sites/default/files/value-text-mining.pdf

Kroes, Neelie (2012). "From crisis of trust to open governing". *European Commission. Press release database*. Bratislava, 5 March.

http://europa.eu/rapid/press-release_SPEECH-12-149_en.htm

Monastersky, Richard (2013). "Publishing frontiers: The library reboot". *Nature*, 27 March. *http://www.nature.com/news/publishing-frontiers-the-library-reboot-1.12664*

Murray-Rust, Peter; **Molloy, Jennifer C.**; **Cabell, Diane** (2014). "Open content mining". En: Moore, Samuel (ed.). *Issues in open research data*, Ubiquity Press, pp. 11-31. ISBN: 978 1909188327

Nualart, Jaume; **Pérez-Montoro, Mario**; **Whitelaw, Mitchell** (2014). "How we draw texts, a review of approaches to text visualization and exploration". *El profesional de la información*, v. 23, n. 3, pp. 221-235. *http://dx.doi.org/10.3145/epi.2014.may.02*

Rosling, Hans (2010). *200 countries, 200 years, 4 minutes - The joy of stats - BBC Four*. *https://www.youtube.com/watch?v=jbkSRLYSojo*

Rowley, Jennifer (2007). "The wisdom hierarchy: representations of the DIKW hierarchy". *Journal of information science*, v. 33, n. 2, pp. 163–180. *http://dx.doi.org/10.1177/0165551506070706*

Serrano-Cobos, Jorge (2014). "*Big data* y analítica web. Estudiar las corrientes y pescar en un océano de datos". *El profesional de la información*, v. 23, n. 6, pp. 561-565. *http://www.elprofesionaldelainformacion.com/contenidos/2014/nov/01.pdf http://dx.doi.org/10.3145/epi.2014.nov.01*

Torres-Salinas, Daniel (2014). "Una visión bibliométrica y dinámica de la investigación biomédica en España". *Indicadores en ciencia y tecnología*. *http://goo.gl/Z1BSH3*

Torres-Salinas, Daniel (2011). "Integrados en la investigación: los *embedded librarians*". *Anuario ThinkEPI*, v. 5, pp. 48-51.

Luis-Millán González-Moreno
Universitat de València

Fernanda Peset
Universitat Politécnica de València
mpeset@upv.es

* * *

Bibliotecarios-documentalistas: hay que aprender a gestionar datos

Estefanía Aguilar-Moreno

Los bibliotecarios/documentalistas o como queramos llamarnos, no deberíamos perder el tren de los datos como motor de la economía que la UE predica. Es necesario desarrollar las competencias aplicadas al sector de los datos que un bibliotecario debería desempeñar, bien desde la formación reglada (lo más deseable) u otro tipo de formación, incluyendo el autoaprendizaje. En esta línea **Eva Ortoll** (*UOC*) y yo publicamos un post recientemente en *ComeIN* (**Aguilar-Moreno**; **Ortoll**, 2015).

Por otro lado me gusta ver que se utilizan mapas y herramientas como *CartoDB*, como forma de visualización de datos, y se apuntan como herramientas imprescindibles para su explotación. Los sistemas de información geográfica que soportan los datos geográficos (el tipo de dato más abundante después de los estadísticos), como apoyo a la economía de los datos y a la toma de decisiones son hace tiempo herramientas del día a día en bibliotecas estadounidenses y canadienses. Asimismo, la información geográfica como tipo específico de datos (tanto abiertos, como de investigación), con sus particulares características, necesita de un perfil que la gestione, quedando este perfil desierto actualmente en el contexto español y europeo. ¿Podríamos los bibliotecarios adentrarnos en este jardín? Las experiencias americanas son envidiables, y últimamente venimos alertando de la oportunidad que los datos (particularizando en los datos geográficos) suponen para bibliotecas y bibliotecarios. Hablamos de sus bondades, utilidad y posibilidades de ampliación de servicios desde la biblioteca en *Geobibliotecas*.

Un comentario sobre *CartoDB* y *Google Maps*

Para hacer una visualización sencilla de pocas capas de datos, cualquiera de los dos puede ir bien. A mayor complejidad, mayores requerimientos de software, como pasa con todo.

CartoDB se basa en una base de datos espacial *PostGIS*. Han creado una interfaz muy amigable para que el usuario no interactúe directamente con los detalles de la base datos. La importación de datos es muy sencilla así como la generación de mapas a partir de las tablas de la base de datos.

Google Maps, más conocido, ofrece muchas más funciones que *CartoDB* a través de su API. Para un proyecto que necesite programación, como la generacion de mapas personalizados, *Google Maps* tiene más posibilidades. Para un proyecto que muestre una visualización en forma de mapa a partir de unos datos, con poca o nula programación, tal vez *CartoDB* sea más intuitivo y se acople mejor a esa necesidad.

Aguilar-Moreno, Estefanía; **Granell-Canut, Carlos** (2015). *Geobibliotecas*. Barcelona: EPI-UOC. ISBN: 978 84 9064 581 9

Aguilar-Moreno, Estefanía; **Ortoll, Eva** (2015). "Economía de los datos: reto y oportunidad". *ComeIn*, n. 40. *http://goo.gl/iAI2Uo*

Estefania Aguilar-Moreno
Universitat Jaume I
Institute of New Imaging Technologies
esamon66@gmail.com